Bioinformatics, 31(16), 2015, 2748–2750
doi: 10.1093/bioinformatics/btv200
Advance Access Publication Date: 9 April 2015
Applications Note

OXFORD

Sequence analysis

AVIA v2.0: annotation, visualization and impact analysis of genomic variants and genes

Hue Vuong*, Anney Che, Sarangan Ravichandran, Brian T. Luke, Jack R. Collins and Uma S. Mudunuri

Advanced Biomedical Computing Center, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 23, 2014; revised on March 12, 2015; accepted on April 5, 2015

Abstract

Summary: As sequencing becomes cheaper and more widely available, there is a greater need to quickly and effectively analyze large-scale genomic data. While the functionality of AVIA v1.0, whose implementation was based on ANNOVAR, was comparable with other annotation web servers, AVIA v2.0 represents an enhanced web-based server that extends genomic annotations to cell-specific transcripts and protein-level functional annotations. With AVIA's improved interface, users can better visualize their data, perform comprehensive searches and categorize both coding and non-coding variants.

Availability and implementation: AVIA is freely available through the web at http://avia.abcc.ncifcrf.gov.

Contact: Hue.Vuong@fnlcr.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Exome sequencing has become the most popular means to evaluate variants in the context of disease. Many clinical panels focus on specific known genes associated with specific diseases or pathways. Accessing all available data about a given mutation can be difficult; from comprehending each database or algorithm to obtaining the computational power to annotate, filter and prioritize lists of variants. To this end, an updated version of the Annotation, Visualization and Impact Analysis web server (AVIA v.2.0) is presented, which incorporates features to help users understand processes within a cell as it relates to disease. AVIA v2.0 is a unique annotation portal in that it incorporates information from >40 annotation databases and allows for custom annotations. The new features of AVIA v.2.0 increase its applicability as a hub for genomics, gene, and protein annotations.

2 Tool Description

The first version of AVIA (Vuong *et al.*, 2014) leveraged ANNOVAR (Wang *et al.*, 2010) as the foundation for gene-based annotations and focused primarily on SNP-level annotations. AVIA v2.0 departs from its previous version by providing integrated access to genomic and proteomic databases and functional annotations. In addition, the availability of a wider range of input options, including gene lists and protein mutations, classification of potentially significant SNPs and genes, and pathway visualization layered with annotation data, also distinguishes AVIA v2.0 from other annotation servers. A comparison between our server and others is available at http://avia.abcc.ncifcrf.gov/apps/site/compare. Conversion tools are also available to assist in navigating AVIA between genomic and protein tools. The integration of bioDBnet (http://biodbnet.abcc.ncifcrf.gov) (Mudunuri *et al.*, 2009), a conversion tool that easily

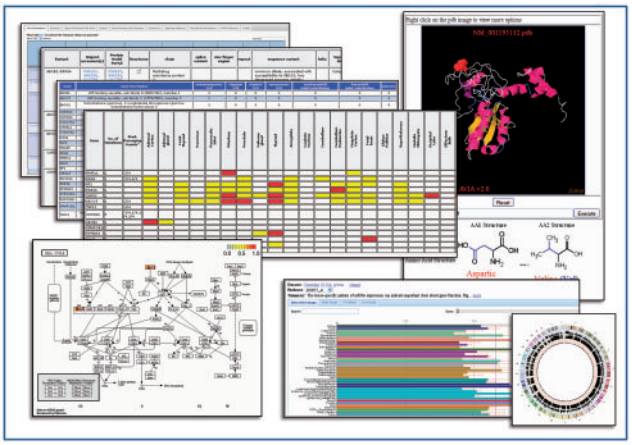

Fig. 1. Expanded views of each tab on AVIA results page

retrieves gene-level annotations using any biological identifier, facilitates the discovery of the relationships between SNPs, genes, expression, protein and biological networks for analysis. As seen in Figure 1, the improved results page is more intuitive, more informative and better suited for exploratory analysis.

2.1 Enhancement of features

AVIA v2.0 improves upon several tools for exploratory analysis that are not offered as standalone web applications elsewhere. A javascript protein structure visualization viewer, JSmol (http://sourceforge.net/projects/jsmol), was incorporated to view users' mutations in the context of protein structure along with a side-by-side comparison of the amino acid properties. Since many full length protein structures are not available within the Protein Data Bank (http://www.rcsb.org), we have begun modeling proteins using I-Tasser (Roy *et al.*, 2010) to attain a comprehensive protein structure library. Our set of full length predicted protein structure is available to shed insight on how an amino acid substitution could potentially affect the protein folding or function. The set of predicted structures is continually being expanded and partial experimental structures are used whenever they contain the position of the observed variant.

AVIAv2.0 helps users classify their variants by automatically flagging those of significant consequence, e.g., protein damaging or clinically relevant variants, as a separate annotation column in the report. In addition, a gene list summarizing variants in genes based on specific categories, e.g., genes with multiple damaging scores the protein scoring algorithm or post-translational modifications, is displayed in a heatmap for each category. AVIA also seamlessly integrates the DAVID-WS API (Jiao *et al.*, 2012) for functional analysis and clustering of genes and FunSeq2 (Fu *et al.*, 2014) for scoring large genomic variant datasets by functionally relevant coding and non-coding variants. The addition of these tools helps prioritize genes for further investigation and associates specific mutations to dysfunction through clinically and functionally relevant assessments.

AVIA v2.0 also adapts PathView (Luo and Brouwer, 2013), an R package that renders biological data on top of the KEGG (Kanehisa *et al.*, 2012) pathway maps, to help users visualize how affected genes may act as hubs within pathways or how their interactions may disrupt cellular activity. For AVIA, 'state' data are shown to reflect the degree to which several databases score damaging genes based on variant data. By default, there are six states showing the scores for the protein scoring databases and prioritization (represented between 0 and 1) for a gene. The most severe score for each database is used to represent that gene (i.e., if there are two mutations for the same gene with different score for FunSeq2, the most severe score is chosen). KEGG pathways containing the users' genes are displayed. Users can also interact with the KEGG pathway map by clicking on their highlighted gene and viewing the additional gene information in KEGG. Future implementations of the PathView feature in AVIA will allow users to select which databases to represent as state data.

For information on the implementation of any of these tools, please see Supplemental Data S1.

2.2 Databases

To our knowledge, AVIA v2.0 combines more databases than any other variant annotation web server and now offers gene-based annotations ranging in categories from regulation and expression data to protein and network annotation. Data obtained from the Gene Expression Atlas (GNF) through the BioGPS portal (Wu *et al.*, 2009) are included to show tissue-specific normal gene expression levels and interaction networks. The expression levels for each gene were ranked by tissue type and then divided into three ranks, representing high, medium and low expressions. Users' data with ranked expression levels for each gene and tissue type are displayed on the interactive results page along with summary information of damaging mutations and integration of BioGPS's graphic API for visualization (http://biogps.org). Normal gene expression may be helpful to researchers performing exploratory work focusing on tissue specificity. Other databases covering post-translational modifications, interactome, metabolome, pharmacogenomics information and clinical data are made available for the users' gene list. Each database will help guide exploratory analysis and may offer insight into genes and their associations in the context of disease.

3 Conclusion

As the number of databases and tools become more abundant, it becomes increasingly time consuming for researchers to leverage all of the available tools and resources. Since very little is known about many genes and their role in the context of disease, our approach is to provide users with the most comprehensive, up-to-date data available. This would promote exploratory analysis and pathway visualization and help highlight biological dysfunction and increase the understanding of cellular processes as they relate to disease.

Together, the extra functionalities detailed here help AVIA v2.0 mature as a hub for genomics, gene and protein annotations by integrating several different types of databases and applications in a fast, comprehensive and significant manner.

Acknowledgements

The authors wish to thank Michele Mehaffey for her invaluable feedback on the application.

Funding

This work was supported with federal funds from the National Cancer Institute, National Institutes of Health [contract HHSN261200800001E]. The content of this publication does not necessarily reflect the views of policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. government.

Conflict of Interest: none declared.

References

Fu,Y. *et al.* (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.

Jiao,X. *et al.* (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics (Oxford, England)*, **28**1805–1806.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Luo,W. and Brouwer,C. (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics (Oxford, England)*, **29**, 1830–1831.

Mudunuri,U. *et al.* (2009) bioDBnet: the biological database network. *Bioinformatics (Oxford, England)*, **25**, 555–556.

Roy,A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Vuong,H. *et al.* (2014) AVIA: an interactive web-server for annotation, visualization and impact analysis of genomic variations. *Bioinformatics (Oxford, England)*, **30**, 1013–1014.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wu,C. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.